Choice Models for Simulating the Consumption of Recommendations

Discussion Paper

Naieme Hazrati*, Francesco Ricci

Free University of Bozen-Bolzano, Bolzano, Italy

Abstract

Recent Recommender Systems (RSs) research has focused on identifying and understanding factors that determine the choice behaviour of their users. By simulating users' choices, influenced by RSs, it was shown that algorithmic biases, such as the tendency to recommend popular items, are transferred to the users' choices. In this position paper, we briefly summarise previous results showing that the effect of an RS on the quality and distribution of the users' choices can be influenced by the users' tendency to prefer certain types of items, i.e., popular, recent, or highly-rated items. To quantify this impact, we have defined alternative Choice Models (CMs) and simulated their effect when users are exposed to recommendations. We found that a bias determined by an RS, e.g., the tendency to concentrate the choices on a restricted number of items, can also be enforced by the CM. Moreover, we have discovered that the quality of the choices can be jeopardised by a CM. We also found that for some RSs, the impact of the CM is less prominent, and their biases are not modified by the CM. This research line shows the importance of assessing algorithmic biases in conjunction with a proper model of users' behaviour.

Keywords

Simulation, Recommender systems, Choice model

1. Introduction

The current analysis of the biases of Recommender Systems (RSs), e,g. the popularity bias, has so far focused on the impact of the RS algorithm and training data on the distribution of the produced recommendations [1]. However, in practice, users are never passively picking the recommended items; they compare them with benchmarks (decision goal), and finally they make a choice. Hence, real users' choices are surely determined by the RS, but also by the users' choice behaviour. Hence, the users' choices overall distribution and quality can be determined by users' tendency to choose items with specific properties, such as, those more popular or recent [2, 3]. For instance, this is clearly observed in how readers purchase books [4].

Therefore, we are interested in understanding the quantitative effects of alternative and "plausible" users' choice behaviours on the distribution and quality of their choices [3, 5]. Aiming at that goal, we operationalise alternative choice models (CMs) that, by mining real purchases data sets, appear to be adopted by real users (e.g., in the Amazon Apps and Games ratings data sets). Then, we use these CMs to simulate users repeatedly choosing items during a

IIR2022: 12th Italian Information Retrieval Workshop, June 29 - June 30th, 2022, Milan, Italy

*Corresponding author.

nhazrati@unibz.it (N. Hazrati*); fricci@unibz.it (F. Ricci)

long time span, among those recommended by an RS [3]. The CMs that we consider, Popularity-CM, Rating-CM, and Age-CM, are influenced by three item properties that users may consider as criteria for making a choice: item popularity, item rating and item age, which is the time difference between the choice and when the chosen item was first available in the system. In fact, these properties have already been studied in the literature [6, 7, 8] as they often influence users in their decision making process. We then use these CMs in a simulation process where users are exposed to recommendations and are simulated to make choices on the base of one of these CMs. We also define a benchmark CM, Base-CM, where the simulated users always select the top recommended item. Base-CM is used to measure the sole effect of the RS and to differentiate the effect of the RS from that of the user's CM. In our empirical analysis we have found interesting properties of the distribution and quality of the users' choices, hence showing the importance of studying the combined effect of a CM and an RS:

1. *The CM can have a significant impact on the distribution and quality of the users' choices.* For instance, when users tend to choose more popular items (Popularity-CM) the choices become even more concentrated over a small set of items. While choosing newer items (i.e., adopting Age-CM) can lead to more diverse choices but with lower quality.

2. *Some important properties and biases of the RS, how they affect the distribution and quality of the choices, are independent of the CM.* In these cases the RS may have unavoidable effects that are not changed by any CM. For instance, the strong choice concentration effect of non-personalised RSs (they recommend the same items to all the users), is not reduced by any of the considered CMs.

Our research line has the potential to enlighten the not yet analysed effect of the users' CM. We aim at understanding the practical implications of the users' population adopting a particular CM. This is important to anticipate the long term effect of an RS on users' decision making.

2. Simulation of Users' Choices

Our research method is based on simulating repeated choices in monthly time intervals, when these choices are influenced by recommendations. In a timestamped data set of users' choices for items, we observe the choices up to a time point t_0, and use them as initial training input for the RS. We then simulate users' choices among the recommendations, in the successive months, and at the end of each month, we retrain the RS with the simulated choices of that month.

Six alternative RSs are studied in our simulations. 1) Popularity-based Collaborative Filtering (PCF) is a nearest neighbourhood collaborative filtering (CF) RS that suggests the most popular items among the choices of nearest neighbour users [9]. 2) Low Popularity-based Collaborative Filtering ($LPCF$) is similar to PCF, but it penalises the ranking score of popular items by multiplying it with the inverse of their popularity. 3) Factor Model (FM) is a CF RS based on matrix factorization [10]. 4) Neural network-based Collaborative Filtering (NCF) leverages a multi-layer perceptron to learn the user-item interaction function that is used to recommend top-k items to the target user [11]. 5) Popularity-based (POP) is a non-personalised RS that recommends the same most popular items to all the users. 6) Average Rating (AR) is another non-personalised RS that recommends items with the highest average ratings.

We assume that the simulated user u, when receives a set of recommendations S_u, uses a multinomial-logit CM to make one choice among the recommended items [9]. The probability of the user u to choose the item j is computed as follows:

$$p(u\ chooses\ j) = \frac{e^{v_{uj}}}{\sum_{k \in S_u} e^{v_{uk}}}$$

where v_{uj} is the utility of the item j for the user u. $|S_u|$ is set to 50 in our experiments. We note that items with a larger utility are more likely to be chosen, but users do not necessarily maximise utility. Based on that multinomial-logit model, we consider four alternative CMs that differ in how the utility of a recommended item is assessed by the simulated user.

- **Rating-CM**: the utility of item j for the user u is equal to their rating prediction, $\hat{r}_{uj}$. We use Inverse Propensity Score Matrix Factorization model (IPS-MF) for such a prediction [12]. We note that Rating-CM is motivated by the assumption that RS users prefer items with larger ratings [13, 2, 5].

- **Popularity-CM:** the utility of the item j is equal to: $v_{uj} = k_f * f_j$, where f_j is the item j popularity (at the time of the user choice), i.e., the number of times j has been chosen in n days prior to the simulated choice divided by n (n=90 in our study). This choice behaviour is often observed and has been extensively studied [6, 14]. To have a fair comparison between the considered CMs, k_f is a constant adjusted so that v_{uj} ranges between 1 and 5, which is the default range of utility values for the Rating-CM (five stars rating).

- **Age-CM:** the utility of item j is equal to: $v_{uj} = k_a * (m - a_j)$, where a_j is the age of item j (at the simulated choice time). Age is the time difference between the choice time and the release date of the item j and m is the maximum item age in the entire data set. k_a, as before, adjusts the impact of the item age on the utility. In Age-CM, more recent items have a larger utility, hence they tend to be preferred by the simulated users. Such a choice behaviour has been observed in some domains [8, 15, 16, 17].

- **Base-CM:** the user always selects the top recommended item. To impose this choice, we set the value of $e^{v_{uj}}$ to 1 if j is the first recommended item and 0 otherwise. The analysis of the choices simulated with Base-CM will show the sole effect of the RS.

3. Experimental Analysis

We have used some *Amazon* data sets to conduct simulation experiments, namely, *Apps* and *Games* data sets [18]. They contain timestamped ratings of users for purchased items. The ratings are provided after the purchase and hence, they signal actual choices. We simulate the final ten months of choice data, while previous months' data were used to bootstrap the simulation (RSs initial training data). We have analysed the full set of the simulated choices using two metrics: (a) the *Gini index* of the chosen items [9], where a higher value of Gini represents a lower diversity of these choices; and (b) *Choice's Rating* which is the average predicted rating (IPS-MF predictions) of the choices, which signals the quality of the choices.

Table 1

		Gini index						Choice's Rating					
Data set	CM\RS	PCF	LPCF	FM	NCF	POP	AR	PCF	LPCF	FM	NCF	POP	AR
Apps	Base	0.94	0.77	0.98	0.95	0.99	0.99	3.94	3.68	3.68	3.88	4.18	4.18
	Age	0.93	0.91	0.94	0.95	0.99	0.99	3.80	3.77	3.69	3.84	4.02	4.23
	Popularity	0.98	0.96	0.98	0.98	0.99	0.99	4.02	3.93	3.74	4.01	4.08	4.27
	Rating	0.93	0.90	0.95	0.94	0.99	0.99	4.01	3.96	3.81	4.07	4.23	4.36
Games	Base	0.96	0.92	0.99	0.96	0.99	0.99	4.20	4.01	4.09	4.17	4.39	4.72
	Age	0.94	0.92	0.96	0.95	0.99	0.99	4.13	4.03	4.03	4.11	4.33	4.63
	Popularity	0.97	0.93	0.99	0.97	0.99	0.99	4.21	4.07	4.08	4.24	4.34	4.61
	Rating	0.95	0.91	0.96	0.95	0.99	0.99	4.25	4.12	4.09	4.24	4.45	4.63

Table 1 presents the Gini index and the Choice's Rating calculated over all of the simulated choices for the 24 analysed combinations of CM and RS [3]. Accordingly, we found that:

The CM can have a significant impact on the distribution and quality of the users' choices. Table 1 shows that when users adopt the Popularity-CM, the Gini index is always larger compared to when other CMs are used. When instead the Age-CM is adopted, the Gini index decreases, however, users' satisfaction drops: Choice's Rating is the lowest with the Age-CM. In addition, Age-CM and Rating-CM reduce the concentration bias of FM: when the users adopt one of these CMs, Gini index is lower than when they adopt the Base-CM. This result suggests that the measured bias of an RS in practice, i.e., when the system is actually used, can be lower than that offline estimated (by using the Base-CM). It is also worth noting that LPCF tends to produce smaller values of Choice's Rating and Gini index by its own, i.e., when Base-CM is used. While when users adopt any of the other CMs, the Choice's Rating and the Gini index increases. This facts contribute to indicate that the actual performance of an RS, i.e., measured in a user study, may be rather different from that estimated without taking into account the possibly even complex choice behaviour of the users.

Certain biases of the RS are independent from the CM. It is important to note that certain biases of the RS are so strong that remain visible in the choices, irrespectively of the CM. In fact, one can look again at Table 1 and note that the non personalised RSs (POP and AR) do have a very large Gini index, and it is not influenced by the CM. It is also interesting to note that LPCF, which is an RS that tries explicitly to suggest unpopular items and clearly has the lowest Gini index, still produces not much more diverse choices, compared with the other RSs, when the effect of a CM is considered (Age-CM, Popularity-CM and Rating-CM).

4. Conclusion

In this position paper we have illustrated a research line that the authors are now conducting: by using a properly defined simulation approach, we measure the effect of alternative users' choice behaviours in the presence of an RS. We are interested in analysing the combined effect of the CM and the RS on the diversity and quality of the choices. We believe that our study can contribute to the start of a new line of research where alternative decision making approaches, potentially followed by the users, are considered in assessing the impact of RS technologies.

References

[1] A. Gunawardana, G. Shani, A survey of accuracy evaluation metrics of recommendation tasks, Journal of Machine Learning Research 10 (2009) 2935–2962.

[2] Z. Szlávik, W. Kowalczyk, M. Schut, Diversity measurement of recommender systems under different user choice models, in: Fifth International AAAI Conference on Weblogs and Social Media, 2011.

[3] N. Hazrati, F. Ricci, The impact of recommender systems and choice models on users' choices, in: International Workshop on Algorithmic Bias in Search and Recommendation, Springer, 2022.

[4] D. W. Heck, L. Seiling, A. Bröder, The love of large numbers revisited: A coherence model of the popularity bias, Cognition 195 (2020) 104069.

[5] N. Hazrati, F. Ricci, Recommender systems effect on the evolution of users' choices distribution, Information Processing & Management 59 (2022) 102766.

[6] H. Abdollahpouri, G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, L. Pizzato, Beyond personalization: Research directions in multistakeholder recommendation, arXiv preprint arXiv:1905.01986 (2019).

[7] D. Kowald, M. Schedl, E. Lex, The unfairness of popularity bias in music recommendation: A reproducibility study, in: European conference on information retrieval, Springer, 2020, pp. 35–42.

[8] J. Bartels, M. J. Reinders, Consumer innovativeness and its correlates: A propositional inventory for future research, Journal of Business Research 64 (2011) 601–609.

[9] D. Fleder, K. Hosanagar, Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity, Management science 55 (2009) 697–712.

[10] Y. Hu, Y. Koren, C. Volinsky, Collaborative filtering for implicit feedback datasets, in: 2008 Eighth IEEE International Conference on Data Mining, Ieee, 2008, pp. 263–272.

[11] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T.-S. Chua, Neural collaborative filtering, in: Proceedings of the 26th international conference on world wide web, 2017, pp. 173–182.

[12] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, T. Joachims, Recommendations as treatments: Debiasing learning and evaluation, arXiv preprint arXiv:1602.05352 (2016).

[13] N. Hazrati, M. Elahi, F. Ricci, Simulating the impact of recommender systems on the evolution of collective users' choices, in: Proceedings of the 31st ACM conference on hypertext and social media, 2020, pp. 207–212.

[14] Z. Zhu, J. Wang, J. Caverlee, Measuring and mitigating item under-recommendation bias in personalized ranking systems, in: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, 2020, pp. 449–458.

[15] F. Zhang, N. J. Yuan, D. Lian, X. Xie, Mining novelty-seeking trait across heterogeneous domains, in: Proceedings of the 23rd international conference on World wide web, 2014, pp. 373–384.

[16] P. Gravino, B. Monechi, V. Loreto, Towards novelty-driven recommender systems, Comptes Rendus Physique 20 (2019) 371–379.

[17] J. Eelen, P. Verlegh, B. Van den Bergh, Exploring the effectiveness of the label "new" in product packaging and advertising, ACR North American Advances (2015).

[18] R. He, J. McAuley, Ups and downs: Modeling the visual evolution of fashion trends with

one-class collaborative filtering, in: proceedings of the 25th international conference on world wide web, 2016, pp. 507–517.